The goal of this worksheet is for you construct a line of best fit to some data points.

The big picture is the following. If the system

$$A\mathbf{z} = \mathbf{b}$$

does not have a solution, because $\mathbf{b}$ does not lie in the column space of $\mathbf{b}$, you can solve instead the *normal* equations

$$A^t A\mathbf{z} = A^t\mathbf{b}.$$

This system will always have a solution, and the solution will be the point $\mathbf{z}$ in the column space of A such that $A\mathbf{z}$ is as close to $\mathbf{b}$ as possible, in the sense that the length

$$||A\mathbf{z} - \mathbf{b}||$$

is minimized.

We want to fit a line to the following (x, y) pairs.

$$(1, 3/2), (2, 3), (3, 0), (4, 2)$$

Yes, there is a fraction. Bummer.

1. Make a sketch, by hand or using Matlab, to visualize the data set.

2. Set up, longhand, equations to solve for m and b to find a line $y = mx + b$ that passes through each of these data points.

3. The equation from the previous step can be written in the form

$$A\mathbf{z} = \mathbf{b}$$

where $\mathbf{z} = (m, b)$. What is the matrix A? What is the vector $\mathbf{b}$? (I.e., concretely write down what these object are in terms of actual numbers)

4. Explain why, just glancing at A, that you do not expect there to be a solution.

5. Find a basis for the left-null space of A and use it to verify that

$$A\mathbf{z} = \mathbf{b}$$

does not have a solution.

6. Instead, we will find a best fit in the following sense. Given a line $y = mx + b$, it generates four data points at our four x-coordinates:

$$\hat{y}_k = mx_k + b$$

where $(x_1, x_2, x_3, x_4) = (1, 2, 3, 4)$. Let $(\bar{y}_1, \bar{y}_2, \bar{y}_3, \bar{y}_4) = (3/2, 3, 0, 2)$. We want to minimize the error between $\bar{\mathbf{y}}$ that comes from our original data and $\hat{\mathbf{y}}$ that comes from the line, in the sense that we want to minimize

$$E = ||\hat{\mathbf{y}} - \bar{\mathbf{y}}||.$$

Rewrite this quantity so that it involves the matrix A and the unknown vector $\mathbf{z} = (m, b)$.

7. Sketch, by hand, the lines corresponding to the following choices of (m, b): $(0, 0)$, $(0, 3)$, $(1, 0)$ and $(0, 2)$. Which of these four lines do you think has the smallest value of E? Then compute E for each of these cases.

8. Set up a linear equation to solve for a best fit (m, b).

9. Now solve it and see if it gives a reasonable answer.

10. Challenge! Go back to your answer to problem 5. Each basis vector gives you a condition that $\mathbf{b}$ must statisfy in order for there to be a solution of $A\mathbf{z} = \mathbf{b}$. Explain, in terms of geometry, slopes, rises, runs or similar what these two conditions actually are.